NASA CONTRACTOR REPORT 166457

Comparative Evaluation of Twenty Pilot Workload Assessment Measures Using A Psychomotor Task In A Moving Base Aircraft Simulator

Sidney A. Connor
Walter W. Wierwille
Virginia Polytechnic Institute & State Univ.
Blacksburg, Virginia 24061

prepared for
Ames Research Center
Under Contract NAG 2-17

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California 94035

NASA CONTRACTOR REPORT 166457

Comparative Evaluation of Twenty Pilot Workload Assessment Measures Using A Psychomotor Task In A Moving Base Aircraft Simulator

Sidney A. Connor
Walter W. Wierwille

CONTRACT NAG 2-17
January 1983

NASA

SUMMARY

A comparison of the sensitivity and intrusion of twenty pilot workload assessment techniques was conducted using a psychomotor loading task in a three degree of freedom moving base aircraft simulator. The twenty techniques included opinion measures, spare mental capacity measures, physiological measures, eye behavior measures, and primary task performance measures. The primary task was an instrument landing system (ILS) approach and landing. All measures were recorded between the outer marker and the middle marker on the approach. Three levels (low, medium, and high) of psychomotor load were obtained by the combined manipulation of windgust disturbance level and simulated aircraft pitch stability. Six instrument rated pilots participated in four sessions lasting approximately three hours each.

Two opinion measures, one spare mental capacity measure, one physiological measure, and one primary task measure demonstrated sensitivity to psychomotor load in this experiment. These measures were Cooper-Harper ratings, WCI/TE ratings, time estimation standard deviation, pulse rate mean, and control movements per unit time. The Cooper-Harper ratings, WCI/TE ratings, and control movements demonstrated sensitivity to all levels of load, whereas the time estimation measure and pulse rate mean only showed sensitivity to some load levels. No intrusion was found for the physiological measures or for the spare mental capacity measures.

The results of this experiment demonstrate that sensitivities of workload estimation techniques vary widely, and that only a few techniques appear to be sensitive to psychomotor load.

INTRODUCTION

The increasing complexity of aircraft systems and the changing roles of pilots and other aircrew personnel have resulted in the need for techniques to measure operator workload in a wide range of situations and tasks. One need only initiate a preliminary survey of the literature on operator workload assessment techniques to discover that a voluminous mass of information has accumulated rapidly in the past two decades. However, major reviews of this literature have concluded that while workload research has advanced in both scope and technology, basic questions remain to be answered for the practitioner who wishes to select workload measures for a given application (Wierwille and Williges, ref. 1). Hicks and Wierwille (ref. 2) have pointed out that, in particular, the lack of information on the relative sensitivity, the degree of intrusion, and the range of transferability of individual techniques makes it difficult for a practitioner to select a workload estimation technique for a given task.

The purpose of the present study was to help fill the need for practical information. Specifically, techniques for measurement of pilot workload were select-

ed for comparison of their sensitivity to psychomotor load in an instrument approach and landing task. The intrusion of these workload assessment techniques was also examined. Follow-on studies which emphasize other major aspects of operator behavior (i.e., perceptual, mediational, and communications) will aid in determining the transferability of workload measurement techniques.

Table 1 provides a list of the workload measurement techniques selected for inclusion in the present study. These techniques were selected on one of two bases. First, evidence was found which indicated that the measures might be sensitive indicators of pilot workload in both simulated and operational flight. Second, previous research had shown that these measures could be useful in a variety of tasks relevant to the flight environment. A review of the twenty techniques selected can be found in Connor (ref. 3).

METHOD

Subjects

Six male instrument-rated pilots served as subjects in this experiment. Each subject was paid $100 for participation (approximately $10 per hour). Participation in the experiment was voluntary. The flight time of the subjects ranged from 500 to 2700 hours with a mean of 1300 hours. The age of the subjects ranged from 23 to 50 years with a mean of 31 years.

Apparatus

Flight Simulator

The primary apparatus in this experiment (Figures 1 and 2) was a modified flight task simulator (Singer Link, Inc., General Aviation Trainer, GAT-1B). The simulator had three degrees of freedom of motion (roll, pitch, and yaw). Translucent blinders were used to cover the windows of the simulator to reduce outside distractions and cues and to aid in the control of cockpit illumination. The simulator had a microphone and speaker inside the cockpit so that two-way communication was possible between the experimenter and the subject at all times during a session.

Several modifications to the flight simulator were made for the experiment. These modifications permitted primary task load manipulation, secondary task operations, response measurement, and scoring.

Primary task load manipulation was accomplished by changing aircraft pitch stability and random windgust disturbance level simultaneously. A circuit modification in the pitch rate computation was incorporated which allowed the pitch rate gain and predominant time constant to be increased together. The simulator already had a level control for random gust amplitude.

Opinion Measurement Systems

The rating scales used for obtaining the pilot opinion data were the Cooper Harper scale (ref. 4) and the Workload-Compensation-Interference/Technical Effectiveness (WCI/TE) scale (ref. 5). Responses were taken by having the subject circle the rating selected for each run using a red marker-pen.

For the Cooper-Harper scale, the circled rating became the score for the run. For the WCI/TE scale, however, the cell of the matrix selected by the subject was mapped to a new (interval) value according to the procedure provided by Donnell and O'Connor (ref. 6). Thereafter the new value was subtracted from 100 to obtain increasing scores with increasing subjective workload.

Secondary Task Measurement Systems

The secondary tasks used in the present study included digit shadowing, memory scanning, mental arithmetic, and time estimation.

In the digit shadowing task, stimuli were presented on a 0.75-inch high seven-segment LED display mounted in a small box on top of the instrument panel. The box contained a shade which increased the contrast of the display. The display was driven by a random digit generator which had an adjustable digit presentation rate. The generator contained circuits which precluded consecutive presentations of the same digit, thereby making blanking between digits unnecessary. The verbal responses of the subject were obtained by a lapel microphone. These responses were recorded on an audio tape recorder. The score calculated on playback was the percentage of digits presented, but not read, by the subject.

In the memory scanning task, stimuli were presented auditorally over a speaker mounted in the rear of the simulator cabin. Stimuli were recorded on audio tape prior to the experiment and were played during the experiment. One channel reproduced the digits, spoken by a male, at intervals that were 10 seconds apart. The other channel provided an electronic signal to initiate a timer in a specially constructed circuit. Subjects responded to the stimuli by pressing one of the two pushbutton switches mounted on the face of the simulator instrument panel. The digital timer was stopped by a signal from one of these two pushbutton switches. The timer also gave an indication of which pushbutton ("yes" or "no") had been depressed. The digital timer was located at the experimenter's station to permit collection of response times and errors on the memory scanning task. The score calculated was the mean time to respond correctly to the stimulus. Incorrect responses and missing responses were not included in the score.

As in the memory scanning task, verbal stimuli for the mental arithmetic task were presented auditorally over headphones worn by the subject. Male voice stimuli were recorded on audio tape prior to the experiment and were played during the experiment. Verbal responses of the subject were sensed by the lapel microphone. The microphone was connected to a second audio tape recorder and speaker located at the experimenter's station. The score obtained was the percent of incorrect and missing single-digit responses.

In the time estimation task, tape recorded verbal cues, spoken by a male, were used to signal the subject to begin production of a 10-second time interval. The subject signaled the beginning and end of an interval by depressing a microswitch mounted at the left grip of the control yoke. Signals from the microswitch provided start and stop pulses for a digital timer. The digital timer was located at the experimenter's station for data collection. Four measures of time estimation were obtained: the mean, the standard deviation, the absolute error, and the RMS error of the completed time estimates for each run.

Physiological Measurement Systems

The physiological measures recorded in the experiment included pulse rate mean, pulse rate variability, respiration rate, pupil diameter, and voice pattern scores.

Pulse rate was sensed by a plethysmograph which was attached to the antihelix of the subject's ear. The plethysmograph output was connected to a Hewlett-Packard 7807C heart rate monitor and then processed by an EAI-380 hybrid computer for on-line scoring. The scores obtained were the mean and the standard deviation of the instantaneous pulse rate, as outputted by the heart rate monitor. The mean is referred to as the pulse rate mean, and the standard deviation is referred to as the pulse rate variability.

The respiration apparatus consisted of a metal belt worn around the subject's upper abdomen with a transducer located approximately one inch from the subject's abdomen. The human body acts as an antenna for stray 60 Hz noise. When the subject breathes, his abdomen moves relative to the transducer, causing changes in the noise signal at the sensor. The sensor signal was connected to an amplifier system at the EAI-380, detected, and recorded on the Sanborn 350 strip chart recorder as the respiration waveform. Respiration rate was obtained graphically, after completion of all runs. The average number of breath cycles (inhale-exhale representing one cycle) per minute was calculated. When a section of a chart record was uninterpretable, that section was deleted.

For pupil diameter measurement, a Panasonic color video camera Model PK-700 was focused on the subject's eyes. The video camera was mounted outside the windscreen of the aircraft simulator with its zoom lens pointed at the subject's normal head position. An opening was cut in the translucent blinder which covered the windscreen, and a white baffle housed the camera lens so that reflections and visual reference cues were eliminated (Figure 1). The video signals were recorded on a Panasonic video cassette recorder Model NV-8310. A monochrome video monitor was used to examine pupil diameter. The pupil diameter measure was obtained by using the "freeze-frame" mode of the recorder at intervals approximately ten seconds apart. To control for head movements and other undesired influences, the mean ratio of pupil diameter to iris diameter was obtained. It was measured horizontally, only while the subject fixated on the glideslope/localizer indicator.

For the speech pattern analysis of secondary task data, an audio tape recorder was used for recording all verbal responses which occurred during the digit shadowing and mental arithmetic tasks. The recorded verbal responses were played back over a Dektor Inc. Psychological Stress Evaluator, and the resulting vocal signals were scored automatically using a computerized scoring system developed at the Patuxent River Naval Air Test Center, Systems Engineering Test Directorate (ref. 7). The system computes a score based on nonuniformity of microtremor in the utterances. Each flight was divided into five 30 second intervals beginning at the outer marker. The mean of the voice stress scores for utterances occurring in each 30 second segment was then computed. Subsequently, a grand mean based on the five, was computed. The process of taking a grand mean for five segments was used to correct for uneven distribution of utterances over the data interval of the flight.

Eye Behavior Measurement Systems

Two eye behavior measures were obtained from the video tape recording and playback system used for the pupil diameter measures. The eye transition frequency was obtained by computing the average number of shifts in eye fixation from one instrument to another per minute. Similarly eye blink frequency was obtained by computing the average number of blinks per minute.

Primary Task Measurement Systems

The primary task performance measures for this experiment included rms angular localizer error (degrees), rms angular glide slope error (degrees), and control movements per unit time. A control movement was defined as the average count per second of the number of inputs to the flight controls (ailerons, elevator, and rudder). A movement occurred when a control rate reached a velocity of more than 4 percent of full range per second, after having passed through zero velocity. All of the primary task measures were computed using the EAI-380 hybrid computer.

Experimental Design

Sensitivity Analysis Design

A complete 3 x 20 within-subject design was used for the sensitivity analysis. Load was the factor with three levels. Measurement technique (Table 1) was the factor with twenty levels. The experimental design matrix for the sensitivity analysis is presented in Figure 3.

The three levels of load were obtained by the combined manipulation of wind gust disturbance level and simulated aircraft pitch stability. The combinations of these variables at the three levels of load are shown in Table 2. The order of presentation of load levels was completely counterbalanced across subjects and within techniques; i.e., each order of load level was randomly assigned to subjects in each technique so that all orders of load level were presented for each measurement technique. The order of presentation of twenty measurement techniques was partially counterbalanced across subjects using a Latin Square design to control for order effects.

Workload measures from different techniques were taken simultaneously on some of the data collection runs. Only those measures which were not likely to affect each other were taken simultaneously. Table 3 shows the scheme used for combining different measurement techniques for data collection. The combination of measurement techniques shown in the table was, to an extent, based on prevous investigations of workload. Hicks and Wierwille's (ref. 2) study supported the combination in condition 2. The two rating scales were administered in separate measurement conditions in counterbalanced order to prevent the ratings on one scale from biasing the ratings on the other scale. The secondary task measures were divided among several conditions because of potential intrusion and interference. Vocal measures were recorded from the two secondary tasks which required a verbal response as per Schiflett and Loikith's (ref. 7) recommendation.

It should be noted that primary task measures were recorded on all subjects and on all data collection flights for the intrusion analysis. However, only data from measurement condition 1 were used for the sensitivity analysis of the primary task measures.

Intrusion Analysis Design

The intrusion analysis was designed to examine the effect of measurement condition, and the interaction of measurement condition with load on primary task performance. Data for all primary task measures were therefore collected for each flight performed in the six measurement conditions. The experimental design matrix for the intrusion analysis is shown in Figure 4.

GENERAL PROCEDURE

Familiarization Session

Upon arrival, each subject read a description of the experiment and was given a consent form. After signing the consent form the subject read detailed instructions which described the flight task. The subject then flew nine familiarization flights in the aircraft simulator. The same instrument landing approach was used in these flights as in the data collection flights.

The major purpose of this session was to familiarize the subject with the range of flight conditions used for the flight task in the data collection sessions. This was accomplished by using combinations of the pitch stability and wind gust disturbance levels shown in Table 4. The flight conditions and the order of these conditions in the familiarization session were the same for all subjects. Subjects were informed of the crosswind conditions. Heading correction instructions were given prior to each flight.

Data Collection Sessions

Each subject participated in three data collection sessions. Each data collection session was designed to obtain data for two of the six measurement conditions listed in Table 3. The general procedures for each measurement condition included three basic elements:

1. measurement condition instructions,
2. one practice flight, and
3. three data collection flights.

These elements are described below.

Instructions. Prior to the practice flight and data collection flights, subjects were given instructions pertinent to their measurement condition. In measurement conditions 1 and 2 subjects read instructions for the rating scale they would be using. In condition 2, subjects read instructions explaining the use of the physiological sensors. In conditions 3, 4, 5, and 6 subjects read instructions concerning the secondary task. The instructions on a secondary task were followed by two sets of practice trials on the secondary task alone to familiarize subjects with the task procedures. Subjects in conditions 3, 4, and 5 were also instructed to perform the secondary task only insofar as this was possible while maintaining adequate performance on the primary task. Adequate performance on the primary task was defined as maintaining localizer and glide slope errors within the "bullseye" of the ILS indicator while holding airspeed constant at 80 mph (+/- 5 mph) between the outer marker and the middle marker. Questions concerning the instructions were then answered, and subjects flew the practice flight.

Practice Flight. The load and crosswind condition in the practice flight was the same as the load and crosswind condition in the first data collection flight. During the practice flight subjects followed the same procedures that they followed in the data collection flights. After the practice flight, questions concerning the procedures were answered. The simulator was then reset to the takeoff location, and the data collection flights were flown.

Data Collection Flights. Three data collection flights were flown by each subject in each measurement condition. All data for the sensitivity and intrusion analyses were collected in these three flights.

The three data collection flights included one flight at each load level. Steady crosswind conditions were used on each flight to disguise the load level. These crosswind conditions were chosen randomly from the list shown in Table 5, without replacement.

During each flight, subjects in measurement conditions 3 to 6 performed a secondary task while flying the ILS approach. All secondary tasks were started approximately one mile before the subject passed over the outer marker. This procedure permitted the subject to become accustomed to performing the secondary task along with the primary task prior to data collection. The secondary tasks were stopped just after the subject passed over the middle marker. Immediately following each flight, subjects in conditions 1 and 2 provided opinion ratings on the appropriate rating scales for the portion of the flight between the outer and middle markers.

Flight Task Procedures

The flight task in this experiment was an Instrument Landing System (ILS) approach to the Seaport Beach runway (29L) which is instrumented in the Singer Link GAT-1B aircraft simulator. Prior to the beginning of a flight, the simulated aircraft was positioned 5 miles outbound from the Seaport Beach outer marker on the 108 degree radial, heading into the wind. When ready to begin, the experimenter informed the subject of the wind direction and speed, and gave him a heading correction for the crosswind. When contacted by the experimenter, the subject took off and climbed to 2000 feet. The subject then flew directly to the outer marker by following the localizer at 100 miles per hour until the glide slope was intercepted. Upon interception of the glide slope, the subject reduced airspeed to 80 miles per hour and proceeded down the glide slope while following the localizer to a landing.

Secondary Task Procedures

Digit Shadowing

For this task the random number generator was set at the maximum reading rate for each subject. This rate was determined prior to the practice flight by increasing the rate over several 30-second trials until omissions began occurring in the subject's responses. Subjects were instructed to read the digits verbally whenever they had free time available from the primary task. Subjects were further instructed to read as many digits as possible, as often as possible, without sacrificing performance on the flight task.

Memory Scanning

The memory scanning task in this experiment is similar to the item recognition task described by Sternberg (ref. 8). Prior to the data collection flights, subjects memorized a set of five single random digits. Subjects were instructed to indicate whether or not a digit was a member of the memorized set by depressing the appropriate response button. Subjects were also instructed to respond as quickly as possible without making errors. Depressing the "yes" response button indicated that the stimulus digit was a member of the memorized set. Depressing the "no" response button indicated that the digit was not a member of the memorized set. Trials were separated by a period of 10 seconds.

Mental Arithmetic

The mental arithmetic procedure used for the present study was similar to the procedure used by Kahneman, Tursky, Shapiro, and Crider (ref. 9). Each trial lasted approximately 10 seconds, with a two-second separation between trials.

On the first second of each trial, the word "ready" was presented over the pilot's headphones. Then, three separate digits were presented on the next three seconds. These digits were randomized such that each digit occurred an equal number of times in each position (i.e., first, second, or third). The subject was instructed to listen carefully to these digits so that he could recall them for the mental arithmetic task. An addition rule (e.g., add 3) followed the presentation of the three digits. The addition rule was followed by a one-second pause. Finally, the word "now" was presented three times over the headphones. The "now's" were presented one second apart. Following each "now", the subject was supposed to speak one of the transformed digits. Subjects were instructed to speak the transformed digits in the order that the original digits were presented.

Time Estimation

The procedure used for time estimation was similar to that recommended by Hart (refs. 10, 11). The subject produced estimated intervals of 10 seconds when signaled to do so. A recorded cue that a trial was about to begin was signaled by the word "ready". This word was followed by a two-second pause and then the word "now". When the subject heard the word "now", he was supposed to designate the beginning of the 10-second interval by depressing the microswitch mounted on the control yoke using his left thumb. When the subject felt that 10 seconds had elapsed from the time he initially depressed the microswitch, he was supposed to indicate the end of the 10 second interval by depressing the microswitch a second time. Subjects were instructed not to count or tap during the time estimation trials. Each trial was separated by a 20 second interval.

RESULTS

Sensitivity Analysis

The complete sensitivity analysis was divided into two major parts:

1. overall sensitivity, and
2. relative sensitivity.

The basic procedure followed in the sensitivity analysis is summarized in Figure 5.

Raw Scores Computations

The data recorded for the different workload measurement techniques in this experiment generally consisted of unscaled signal values or sets of observations taken during a given flight. These data were transformed into appropriate score values for data analysis.

Conversion to z-Scores

Following the raw score computations, the scores from the three data collection flights for each technique were transformed to z-scores (normalized scores). The z-score transformations were performed to prevent differences in scale value (e.g., breaths per minutes vs. localizer rms angular position error) from masking true differences in the techniques.

Overall ANOVA

An overall analysis of variance was performed on the normalized scores to determine whether or not the manipulation of load was effective in the present experiment. Since z-scores were used, a technique main effect was not possible. A significant main effect of load was found, F (2,10) = 5.34, $p < 0.0001$, and a significant load by technique interaction was found, F (38, 190) = 2.76, $p < 0.05$. These results are summarized in Table 6.

Individual ANOVAs

The significant load by technique interaction indicated that the measurement techniques were differentially affected by load. Individual ANOVAs were used to isolate those techniques which were sensitive to load.

The load by technique interaction would normally be investigated by using simple effects F-tests to determine the effect of load on each technique. These tests use the load by technique by subject interaction from the overall ANOVA as a denominator term in the F-ratios. This implicitly assumes homogeneity of variance among the different techniques.

The homogeneity of variance assumption was examined using Bartlett's test for homogeneity of variance (ref. 12). The results of Bartlett's test showed that there was significant heterogeneity of variance among the different techniques, χ^2 (19) = 58.008, $p < 0.01$. Due to the violation of the assumption of homogeneity of variance, the use of simple effects F-tests to examine the load by technique interaction was considered inappropriate.

Further study of the load by technique interaction was conducted using individual ANOVAs on each technique (workload measure) versus load (Table 1). For the purpose of clarity, the results of the ANOVAs are presented in logical groupings of measurement techniques.

The results of the individual ANOVAs performed on the opinion measures are presented in Table 7. The Cooper-Harper scale showed a significant effect of load, F (2, 10) = 16.39, $p = 0.0007$, and the WCI/TE scale also showed a significant effect of load, F (2,10) = 31.15, $p < 0.0001$. The mean normalized scores for these two rating scales are plotted as a function of load in Figure 6 for the Cooper-Harper scale and in Figure 7 for the WCI/TE scale.

The results of the individual ANOVAs performed on the spare mental capacity measures are presented in Tables 8 and 9. None of the measures in Table 8 showed a significant effect of load. Among the time estimation measures presented in Table 9, standard deviation alone showed a significant load effect, F (2, 10) = 5.69, p = 0.022. The mean normalized scores for the time estimation standard deviation measure are plotted in Figure 8.

The results of the individual ANOVAs performed on the physiological measures are summarized in Table 10. Only pulse rate mean showed a significant effect of load, F (2, 10) = 8.89, p = 0.006. The normalized means for the pulse rate mean measure are plotted as a function of load in Figure 9.

The results of the individual ANOVAs performed on the eye behavior measures are summarized in Table 11. Neither of these measures showed a significant effect of load.

The results of the individual ANOVAs performed on the primary task measures are summarized in Table 12. Only the control movements measure showed a significant effect of load, F (2, 10) = 33.84, $p < 0.0001$. Figure 10 shows the mean normalized scores for the control movements measure plotted as a function of load.

Newman-Keuls Comparisons

The locus and direction of the effect of load on the sensitive techniques was examined using Newman-Keuls comparisons of mean normalized scores between load levels. The results of these comparisons are presented in Table 13.

An examination of Figures 6 and 7 indicates that the mean normalized scores for both rating scales increased monotonically with load. The Newman-Keuls comparisons performed on these data showed that the differences between all pairs of means were statistically significant for both rating scales.

In contrast to the results of the two opinion measures, Figure 8 shows that mean normalized scores for the time estimation standard deviation measure is not a monotonic function of load. The Newman-Keuls comparisons performed on the scores for this measure showed that scores decreased significantly from the low to the medium load condition, and then increased significantly from the medium to the high load condition. The difference between the scores for the low and high load conditions was not significant, however.

The only physiological measure which showed a significant effect of load was pulse rate mean. Scores on this measure increased monotonically with load as shown in Figure 9. However, Newman-Keuls comparisons of the mean normalized scores for the three load conditions showed that the only significant difference in mean normalized scores was between the high and the low load conditions.

Finally, mean normalized scores on the control movements measure increased monotonically with load as shown in Figure 10. Newman-Keuls comparisons for this measure showed that the differences between all pairs of means were statistically significant.

Classification of Techniques

Sensitivity is defined in the present study as the ability of a measure to discriminate statistically between different load conditions. A logical classification of techniques based on demonstrated sensitivity could be generated from an examination of the Newman-Keuls comparisons. This type of classification scheme is presented in Table 14.

Techniques which demonstrated sensitivity to all pairs of load conditions (i.e., low vs. medium, medium vs. high, and low vs. high) were included in class I. These measures are preferred over other techniques which demonstrated only partial sensitivity, or no sensitivity in the present study. Techniques which showed sensitivity to some differences in load conditions (but not all) were included in class II. These measures are less preferred than class I techniques, but are more preferred than class III techniques. Class III techniques did not demonstrate sensitivity to load in the present study. This class includes all techniques except those in class I and class II.

Sample Size Estimates

One possible reason that only five of the twenty workload assessment techniques demonstrated sensitivity in the present study is that the other techniques simply required a greater number of subjects to show a significant effect of load. It is possible to estimate the sample size required to detect a reliable load effect for a given workload assessment technique at specified levels of significance and power. These calculations were performed for those techniques which did not demonstrate sensitivity in the present study, to provide an indication of the practical costs of achieving statistical significance. The procedure used for estimating the sample size required for finding sensitivity is described by Bowker and Liberman (ref. 13). Sample sizes were estimated for a significance level of 0.05 and for a power of approximately 0.80. The results of these estimates are presented in Table 15.

Intrusion Analysis

The equipment and procedures used for some workload assessment techniques may interfere with performance on the primary (flight) task. In the present experiment, data for the twenty measurement techniques were recorded in six measurement conditions as shown in Table 3. These six measurement conditions differed in the equipment and procedures used for data collection. The purpose of the intrusion analysis was to examine the effect of these measurement conditions on primary task performance.

The equipment and procedures used in measurement condition 1 were assumed to be unobtrusive to primary task performance. Primary task performance in this condition was therefore used as a standard of comparison for primary task performance on the other five measurement conditions. The measures of primary task performance

which were used for these comparisons included scores on localizer rms error, glide slope rms error, and control movements per secon

A multivariate analysis of variance (MANOVA) was performed to examine the effect of condition, load, and the interaction of condition and load on the primary task measures. Only the main effect of load was found to be significant F (2, 10) = 9.42, p = 0.0002. These results are summarized in Table 16. Because there was no significant interaction of condition with load, it can be concluded that the physiological measuring equipment and the secondary tasks did not significantly affect pilot performance in terms of the three primary task variables.

CONCLUSIONS

This study has shown that five measures of workload estimation were sensitive indicators of load in a piloting task that is predominantly psychomotor in nature. Another fifteen measures, believed to be "good" measures of workload, showed no reliable effect. The main conclusion that must be drawn from the study is that few measures are sensitive to psychomotor load.

Of the five techniques demonstrating sensitivity, only three exhibited monotonic score increases with load as well as statistically reliable differences between all pairs of load levels. Consequently, only the three meet all criteria for sensitivity to psychomotor load. These class I techniques are the ones that are recommended for measurement of psychomotor load:

Cooper/Harper ratings,
WCI/TE ratings, and
Control movements per second.

The other two techniques showed sensitivity to psychomotor load, but did not discriminate between all pairs of load levels. These class II techniques are:

Time estimation standard deviation, and
Pulse rate mean.

These measures would be helpful in evaluating psychomotor load, but they should not be relied on exclusively. At least one class I technique should also be used in conjunction with these measures.

It is worth noting that only two opinion measures were taken in the present experiment, and both proved sensitive. This suggests that well-designed rating scales are among the best of techniques for evaluating psychomotor load. In regard to the primary task measures, the control movements measure alone was sensitive. However, this measure is also the only primary task measure which reflects "strategy" of the pilot. Consequently, one could speculate that selecting a primary task measure that reflects strategy will most likely result in good sensitivity.

In this experiment, psychomotor load was manipulated by changing random wind-gust disturbance level and aircraft pitch stability simultaneously. The significant main effect of load found in the overall ANOVA for the sensitivity analysis (Table 6) indicates that this was an effective method of varying operator load. Nevertheless, fifteen (techniques) measures showed no reliable change as a function of load. When these fifteen measures are subjected to a power analysis to determine sample size, the number of subjects required ranged from 12 to well over 100 (Table 15). One can only conclude that at best the fifteen measures, as taken, are much less sensitive to psychomotor load than the five appearing in Classes I and II. Of course, there is always the possibility that the measures would be sensitive to loading along other dimensions of human performance, such as psychomotor tasks of a different nature, or mediational or cognitive tasks, for example.

In regard to intrusion, this experiment showed that no significant interference occurred for the physiological measures or for the secondary task measures. Performance as measured using three primary (flight) task measures showed no reliable changes as a function of addition of these measures. These results are in agreement with those obtained by Hicks and Wierwille (ref. 2), who found no significant intrusion for a pulse rate measurement and a digit shadowing task in a driving simulator study. Apparently, although the task of instrument landing is more difficult than driving, pilots are able to maintain performance even though they are called upon to perform additional subsidiary tasks.

In general, the results of the experiment show that there are wide variations in the sensitivity of workload estimation measures. Great care must be taken in selecting measures for a given experiment. Otherwise, it is possible that no changes in workload will be found, when indeed there are changes. The results of the study reported here should be helpful in selecting sensitive measures for psychomotor load. Future studies should be helpful in selecting measures for other dimensions of human behavior.

ACKNOWLEDGEMENTS

The authors wish to thank Mrs. Sandra Hart and Mr. Tom Wempe, NASA Ames Research Center, for their helpful technical suggestions and for serving as grant monitors during this project. The authors also wish to acknowledge the help of Messrs. John G. Casali and Mans Rahimi in equipment construction and experimental design. Finally, the authors thank Dr. Sam Schiflett, Naval Air Test Center, Patuxent River, MD. for making equipment available to perform voice stress measurements.

REFERENCES

1. Wierwille, W. W. and Williges, R. C. Survey and analysis of operator workload assessment techniques. Blacksburg, Virginia: Systemetrics, Inc. Report No. S-78-101, September, 1978.

2. Hicks, T. G. and Wierwille, W. W. Comparison of five mental workload assessment procedures in a moving base driving simulator. Human Factors, 1979, 21, 129-143.

3. Connor, S. A. A comparison of pilot workload assessment techniques using a psychomotor task in a moving base aircraft simulator. Master's Thesis, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, October, 1981.

4. Cooper, G. E. and Harper, R. P., Jr. The use of pilot rating in the evaluation of aircraft handling qualities. Moffett Field, California: National Aeronautics and Space Administration, Ames Research Center, NASA TN-D-5153, April, 1969.

5. Donnell, M. L. The application of decision-analytic techniques to the test and evaluation phase of the acquisition of a major air system: Phase III. McLean, Virginia: Decisions and Designs Technical Report TR 79-6-91, May, 1979.

6. Donnell, M. L. and O'Connor, M. F. The application of decision analytic techniques to the test and evaluation phase of the acquisition of a major air system: Phase II. McLean, Virginia: Decisions and Designs Technical Report TR 78-3-35, April, 1978.

7. Schiflett, S. G. and Loikith, G. J. Voice stress as a measure of operator workload. Patuxent River, Maryland: Naval Air Test Center, Technical Memorandom TM 79-3 SY, December 31, 1979.

8. Sternberg, S. High-speed scanning in human memory. Science, 1966, 153, 652-654.

9. Kahneman, D., Tursky, B., Shapiro, D. and Crider, A. Pupillary, heart rate, and skin resistance changes during a mental task. Journal of Experimental Psychology, 1969, 79, 164-167.

10. Hart, S. G. Time estimation as a secondary task to measure workload-attention sharing effect on operator performance. Proceedings of the Eleventh Conference on Manual Control, NASA-Ames Research Center, May 21-23, 1975, 64-77. (NASA TMX-62, 464).

11. Hart, S. G. Pilot workload during final approach in congested airspace. Proceedings of the 1978 IEEE Conference on Decision and Control, San Diego, California, January 10-12, 1979, 1345-1349.

12. Winer, B. J. Statistical principles in experimental design, (Second edition). New York: McGraw-Hill Book Company, 1971.

13. Bowker, A. H. and Lieberman, G. J. Engineering statistics. New Jersey: Prentice-Hall, Inc., 1959.

TABLE 1

Workload Assessment Techniques Which Were Tested in the Present Experiment

OPINION	
1.	Cooper-Harper Scale
2.	WCI/TE Scale
SPARE MENTAL CAPACITY	
3.	Digit Shadowing (% errors)
4.	Memory Scanning (Mean time)
5.	Mental Arithmetic (% errors)
6.	Time Estimation Mean (Seconds)
7.	Time Estimation Standard Deviation (Seconds)
8.	Time Estimation Absolute Error (Seconds)
9.	Time Estimation RMS error (Seconds)
PHYSIOLOGICAL	
10.	Pulse Rate Mean (Pulses per minute)
11.	Pulse Rate Variability (Pulses per minute)
12.	Respiration Rate (Breath cycles per minute)
13.	Pupil Diameter (Normalized units)
14.	Voice Pattern (Digit Shadowing Task)
15.	Voice Pattern (Mental Arithmetic Task)
EYE BEHAVIOR	
16.	Eye Transition Frequency (Transitions per minute)
17.	Eye Blink Frequency (Blinks per minute)
PRIMARY TASK	
18.	Localizer RMS Angular Position Error (Degrees)
19.	Glide Slope RMS Angular Position Error (Degrees)
20.	Control Movements per second (Aileron + Elevator + Rudder)

TABLE 2

Primary Task Load Conditions

	LOAD CONDITION		
	Low	Medium	High
RANDOM GUST LEVEL	Low	Medium	High
Estimated Std. Dev. (mph)	0	2.7	5.9
PITCH STABILITY	High	Medium	Low
a. Control input to pitch rate output equivalent gain (degrees/s per % of control range)	0.522	3.560	7.83
b. Control input to pitch rate output equivalent time constant (s)	0.097	0.660	1.45

TABLE 3

Combination of Measurement Techniques
for Data Collection

Measurement Condition	Measurement Techniques
1.	Cooper-Harper Scale Pupil Diameter Eye Transition Frequency Eye Blink Frequency Localizer RMS Error Glide Slope RMS Error Control Movements
2.	WCI/TE Scale Pulse Rate Mean Pulse Rate Variability Respiration Rate
3.	Digit Shadowing Voice Pattern
4.	Memory Scanning
5.	Mental Arithmetic Voice Pattern
6.	Time Estimation (Mean) (Std. Dev.) (Abs. Error) (RMS Error)

TABLE 4

Familiarization Flight Pitch Stability, Wind Gust Disturbance, and Crosswind Conditions

Flt. No.	Pitch Stability Level	Wind Gust Disturbance Level	Crosswind Velocity (Knots)	Crosswind Direction (Deg. From)
1.	High	Low	28.5	346
2.	High	Medium	23.5	338
3.	High	High	19.0	327
4.	Medium	Low	16.0	310
5.	Medium	Medium	15.5	277
6.	Medium	High	17.5	257
7.	Low	Low	21.0	243
8.	Low	Medium	26.0	234
9.	Low	High	31.0	227

TABLE 5

Crosswind Conditions Used for the Data Collection Runs

Wind Direction (Degrees From)	Wind Velocity (Knots)
227	31.0
230	28.5
234	26.0
238	23.5
243	21.0
249	19.0
257	17.5
266	16.0
277	15.5
299	15.5
310	16.0
319	17.5
327	19.0
333	21.0
338	23.5
342	26.0
346	28.5
349	31.0

TABLE 6

Summary Table for the
Overall Load x Technique ANOVA

Source	df	SS	F-Ratio
Between Subjects			
Subjects (S)	5	21.8143	
Within-Subject			
Load (L)	2	8.1733	5.34*
L x S	10	7.6514	
Technique (T)	19	0.0000	0.00
T x S	95	207.3533	
Load x Technique (LxT)	38	33.1166	2.76**
L x T x S	190	59.8862	
Total	359	339.9951	

*p < 0.05
**p < 0.0001

TABLE 7

Summary Table for the Rating Scales Individual ANOVAs

Source	df	SS	F-Ratio
	Cooper-Harper Scale		
Subject (S)	5	3.6108	
Load (L)	2	10.2589	16.39*
L x S	10	3.1304	
	WCI/TE Scale		
Subject (S)	5	9.3209	
Load (L)	2	6.6170	31.15**
L x S	10	1.0621	

*p = 0.0007
**p < 0.0001

TABLE 8

Summary Table for the Secondary Task Individual ANOVAs
(Other Than Time Estimation)

Source	df	SS	F-Ratio
	Digit Shadowing (% Errors)		
Subject (S)	5	15.8067	
Load (L)	2	0.3356	1.96
L x S	10	0.8576	
	Memory Scanning (Mean Time)		
Subject (S)	5	13.9333	
Load (L)	2	0.1119	0.19
L X S	10	2.9526	
	Mental Arithmetic (% Errors)		
Subject (S)	5	14.5096	
Load (L)	2	0.5542	1.43
L X S	10	1.9362	

TABLE 9

Summary Table for the Time Estimation Task Individual ANOVAs

Source	df	SS	F-Ratio
	Time Estimation (Mean)		
Subject (S)	5	14.2906	
Load (L)	2	0.2856	0.59
L x S	10	2.4241	
	Time Estimation (Std. Dev.)		
Subject (S)	5	8.5075	
Load (L)	2	4.5176	5.69*
L x S	10	3.9731	
	Time Estimation (Abs. Error)		
Subject (S)	5	11.6827	
Load (L)	2	0.2526	0.25
L x S	10	5.0606	
	Time Estimation (RMS Error)		
Subject (S)	5	9.4391	
Load (L)	2	0.5080	0.36
L x S	10	7.0483	

*p = 0.022

TABLE 10

Summary Table For the Physiological Measures
Individual ANOVAs

Source	df	SS	F-Ratio
Pulse Rate Mean			
Subject (S)	5	16.1136	
Load (L)	2	0.5674	8.89*
L x S	10	0.3191	
Pulse Rate Variability			
Subject (S)	5	15.1790	
Load (L)	2	0.2291	0.72
L x S	10	1.5918	
Respiration Rate			
Subject (S)	5	13.5623	
Load (L)	2	1.2327	2.78
L x S	10	2.2164	
Pupil Diameter			
Subject (S)	5	7.2348	
Load (L)	2	0.0698	0.04
L x S	10	9.9688	
Voice Pattern (Digit Shadowing)			
Subject (S)	5	13.9742	
Load (L)	2	0.5703	1.16
L x S	10	2.4576	
Voice Pattern (Mental Arithmetic)			
Subject (S)	5	10.2929	
Load (L)	2	0.0161	0.01
L x S	10	6.6918	

*p = 0.006

TABLE 11

Summary Table for the Eye Behavior Measures
Individual ANOVAs

Source	df	SS	F-Ratio
	Eye Transition Frequency		
Subject (S)	5	14.7444	
Load (L)	2	0.2889	0.74
L x S	10	1.9608	
	Eye Blink Frequency		
Subject (S)	5	15.3043	
Load (L)	2	0.3625	1.36
L x S	10	1.3336	

TABLE 12

Summary Table for the Primary Task Measures
Individual ANOVAs

Source	df	SS	F-Ratio
	Localizer RMS Error		
Subject (S)	5	13.0838	
Load (L)	2	1.5970	3.43
L x S	10	2.3247	
	Glide Slope RMS Error		
Subject (S)	5	6.8489	
Load (L)	2	1.3508	0.77
L x S	10	8.8005	
	Control Movements		
Subject (S)	5	3.7282	
Load (L)	2	11.5640	33.84*
L x S	10	1.7087	

*$p < 0.0001$

TABLE 13

Results of Newman-Keuls Analyses
on Sensitive Measures*

a. Cooper-Harper Ratings

Load Level	Low	Medium	High
Mean	-0.9389	0.0294	0.9096
	I------I	I------I	I------I

b. WCI/TE Ratings

Load Level	Low	Medium	High
Mean	-0.7848	0.0934	0.6913
	I------I	I------I	I------I

c. Time Estimation S.D.

Load Level	Medium	Low	High
Mean	-0.6556	0.0952	0.5604
	I------I	I----------I	

d. Pulse Rate Mean

Load Level	Low	Medium	High
Mean	-0.2202	0.0057	0.2145
	I----------I		
		I----------I	

e. Control Movements

Load Level	Low	Medium	High
Mean	-0.9830	0.0025	0.9804
	I------I	I------I	I------I

*Means with a common underline do not differ significantly at the $p < 0.05$ level.

TABLE 14

Logical Classification of Techniques
Based on Demonstrated Sensitivity

Class I: Complete Sensitivity Demonstrated
- Cooper-Harper Scale
- WCI/TE Scale
- Control Movements/Unit Time

Class II: Some Sensitivity Demonstrated
- Time Estimation Standard Deviation*
- Pulse Rate Mean **

Class III: Sensitivity Not Demonstrated
- All Other Techniques (See Table 15)

*Double valued function (See text)
**Limited sensitivity (See text)

TABLE 15

Estimated Sample Sizes Required for Achieving a Significant Load Effect for Techniques not Demonstrating Sensitivity

Technique	Estimated Sample Size
SPARE MENTAL CAPACITY	
Digit Shadowing	18
Memory Scanning	>100
Mental Arithmetic	25
Time Estimation (Mean)	53
Time Estimation (Abs. Error)	>100
Time Estimation (RMS Error)	85
PHYSIOLOGICAL	
Pulse Rate Variability	45
Respiration Rate	15
Pupil Diameter	>100
Speech Pattern (D. Shadow.)	28
Speech Pattern (M. Arith.)	>100
EYE BEHAVIOR	
Eye Transition Frequency	42
Eye Blink Frequency	25
PRIMARY TASK	
Localizer RMS Error	12
Glide Slope RMS Error	41

TABLE 16

MANOVA Summary Table for the Intrusion Analysis
(Wilk's Criterion)

Source	df	F-Ratio
Between Subjects		
Subjects (S)	5	
Within-Subject		
Condition (C)	5	0.42
S x C	25	
Load (L)	2	9.42*
S x L	10	
Condition x Load	10	1.00
S x C x L	50	
Total	107	

*$p < 0.0002$

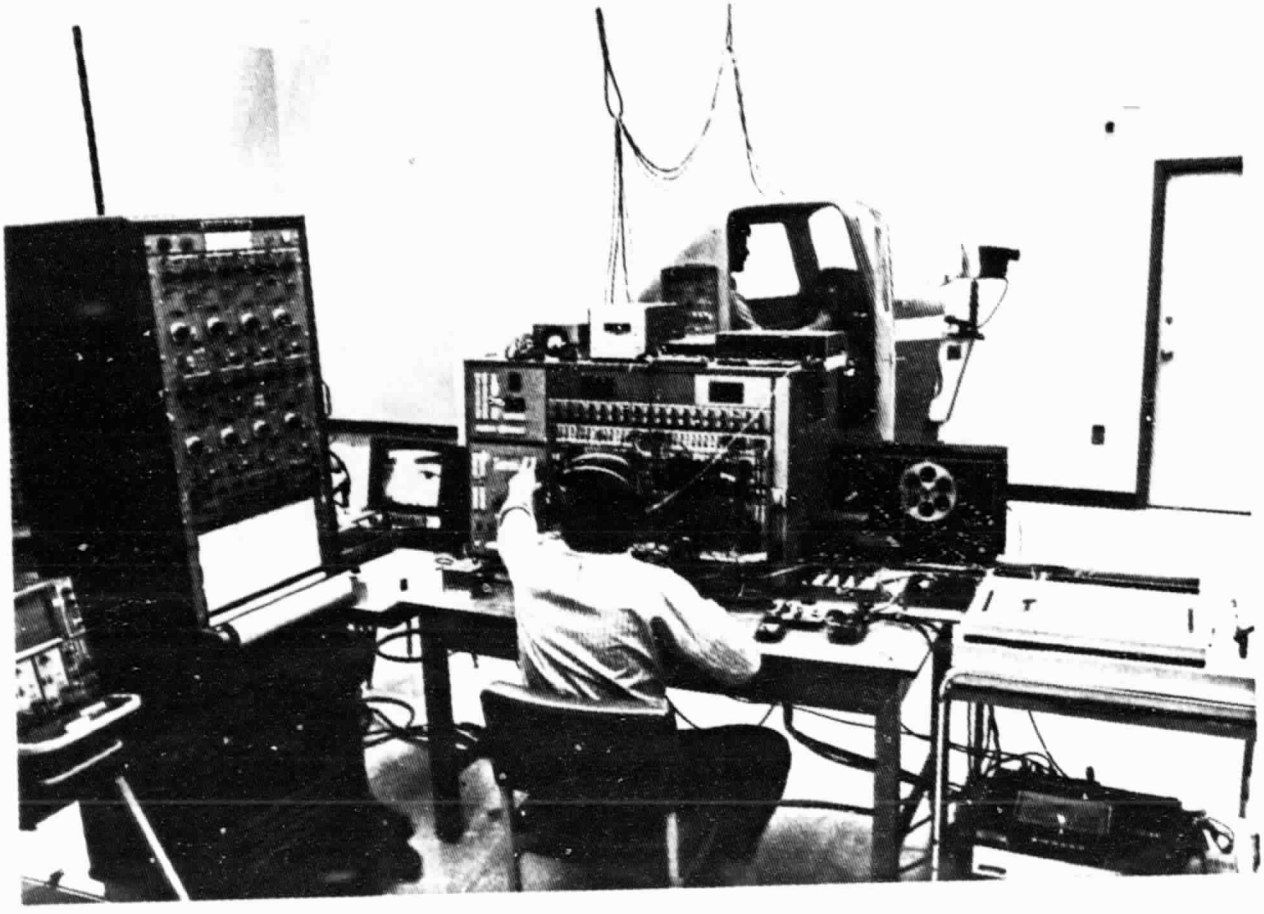

Figure 1. Experimental facility with experimenter's station in the foreground.

Figure 2. Subject wearing physiological sensors in the cockpit of the simulator.

MEASUREMENT TECHNIQUE	LOAD Low	Medium	High
1. Cooper-Harper			
2. WCI/TE Scale			
3. Digit Shadowing			
. . .	. . .	. . .	. . .
20. Control Mov.			

Figure 3. Experimental design matrix for the sensitivity analysis.

MEASUREMENT TECHNIQUE	Low	Medium	High
1. Primary Task Alone			
2. Primary Task with Physiol.			
3. Primary Task with Shadow.			
4. Primary Task with Memory			
5. Primary Task with M. Arith.			
6. Primary Task with Time Est.			

Figure 4. Experimental design matrix for the intrusion analysis.

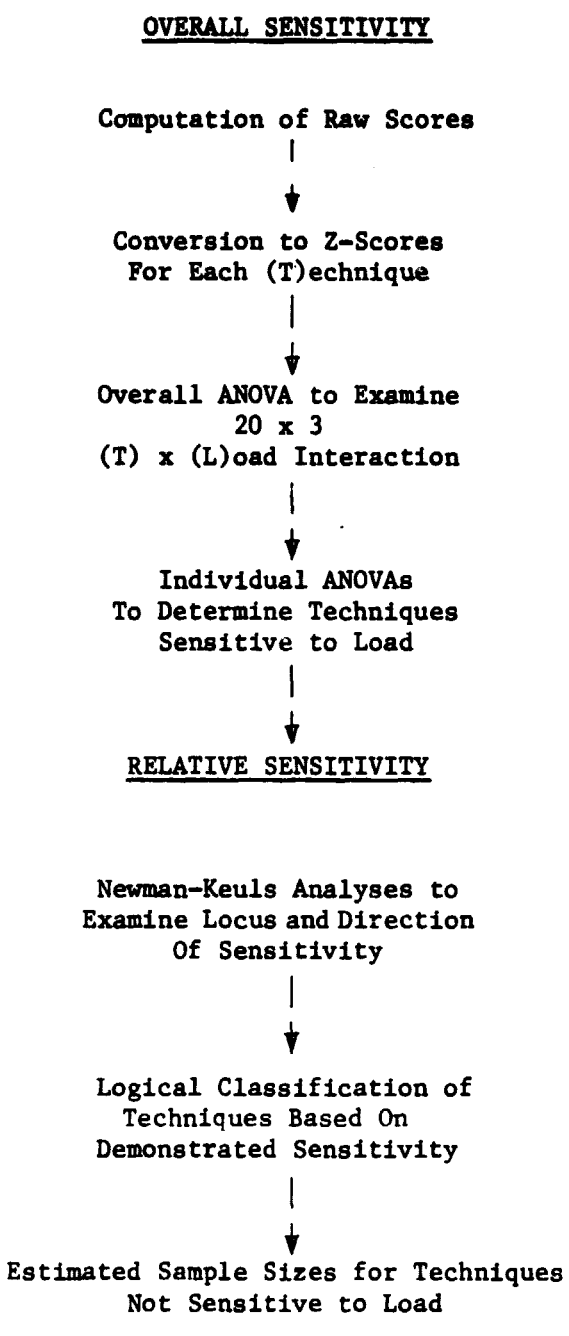

Figure 5. Basic procedure for sensitivity analysis.

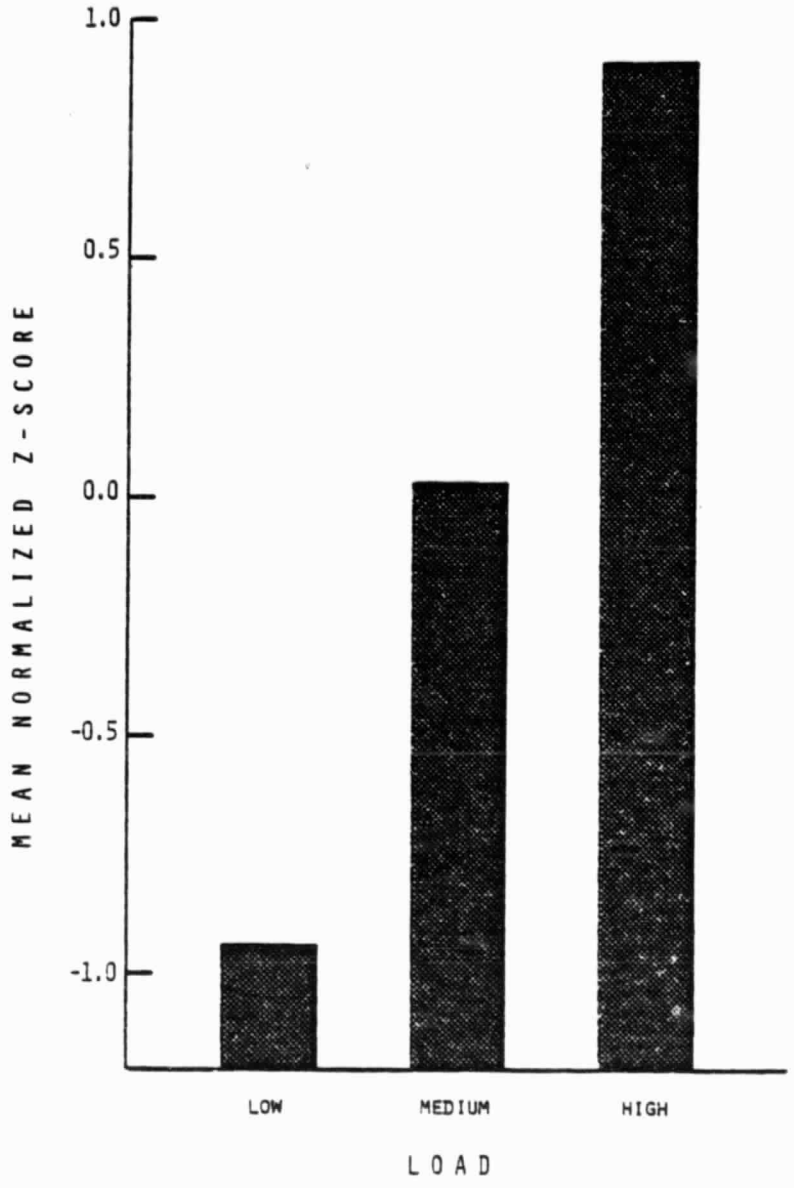

Figure 6. Mean normalized scores plotted as a function of load for the Cooper-Harper rating scale measure.

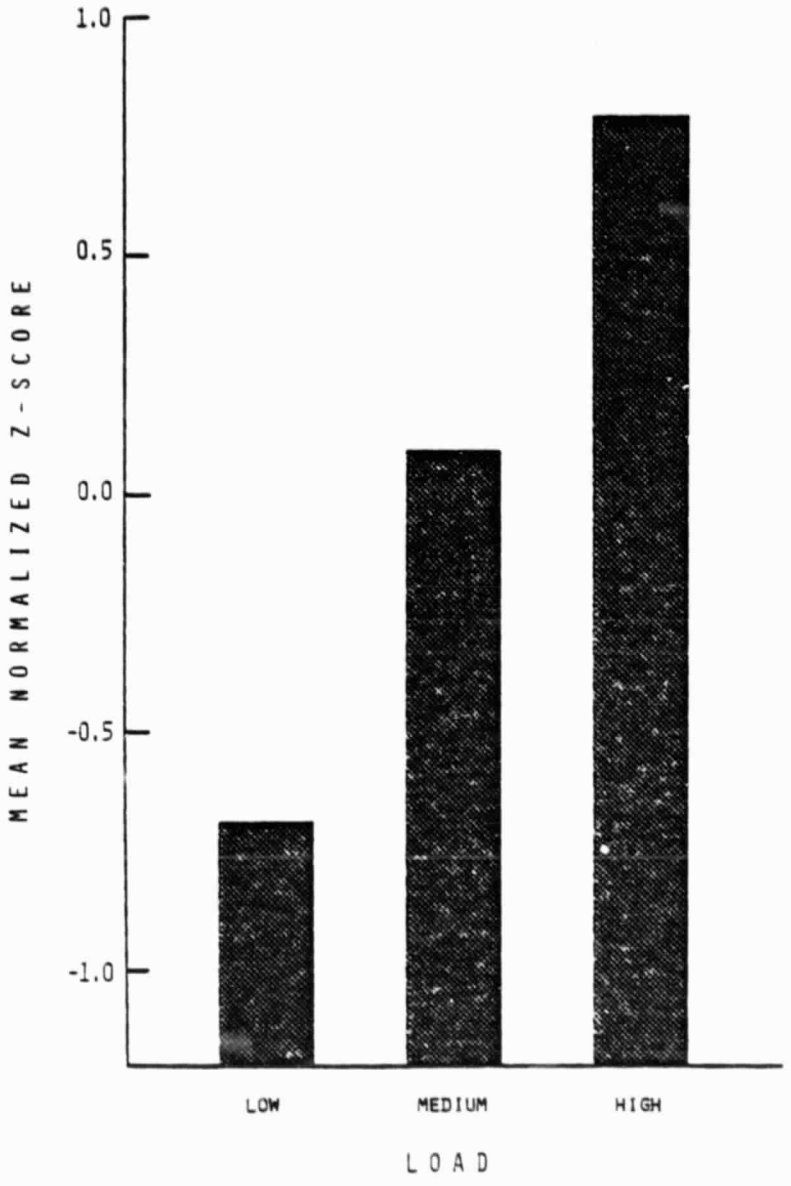

Figure 7. Mean normalized scores plotted as a function of load for the WCI/TE rating scale measure.

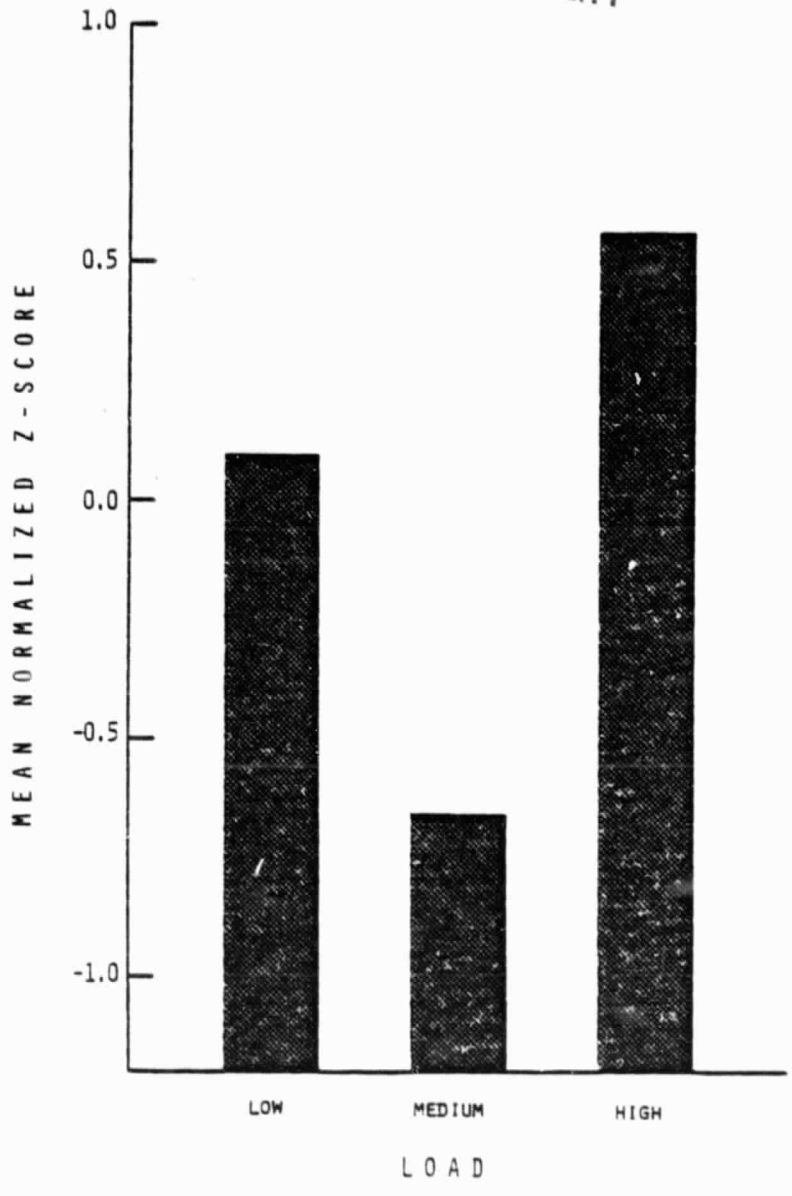

Figure 8. Mean normalized scores plotted as a function of load for the time estimation standard deviation measure.

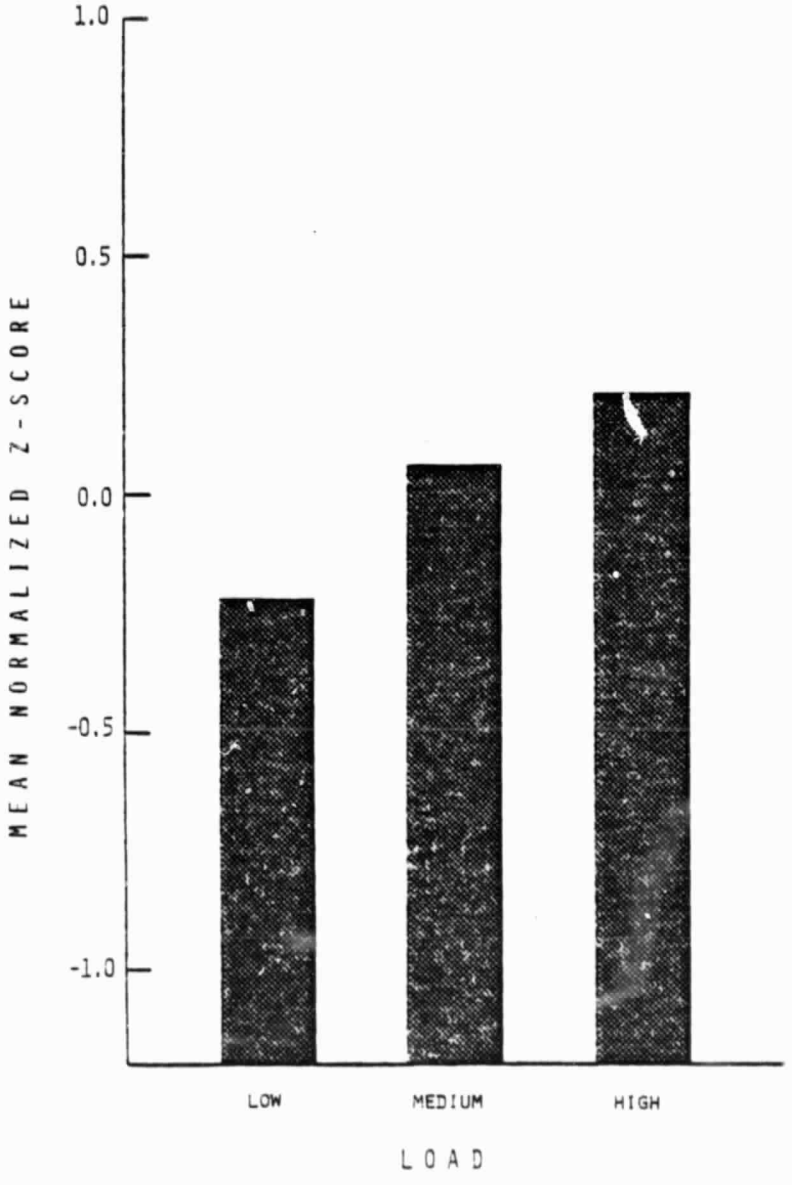

Figure 9. Mean normalized scores plotted as a function of load for the pulse rate mean measure.

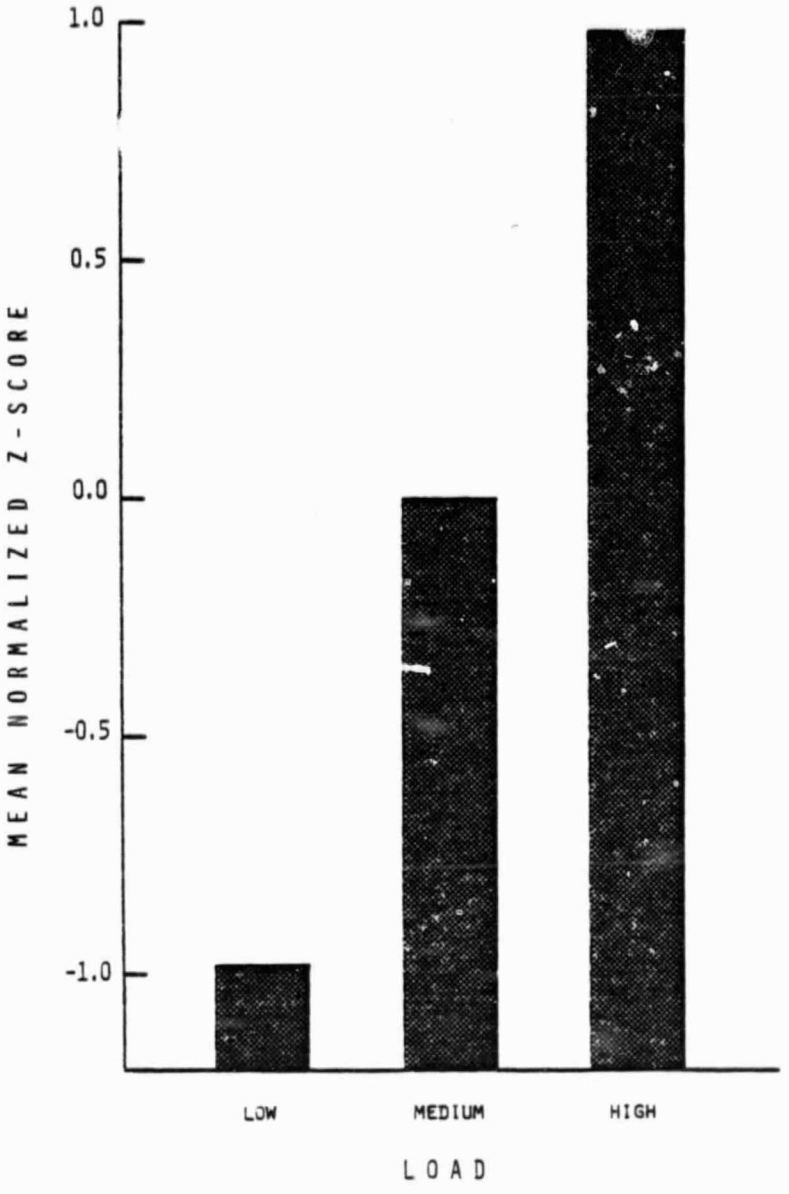

Figure 10: Mean normalized scores plotted as a function of load for the control movements measure.